

A Computational Framework for Nonlinear Dimensionality Reduction of Large Data Sets: The Exploratory Inspection Machine (XIM)

Technical Report

Version 1.0

Axel Wismüller, M.D., Ph.D.

Director, Computational Radiology Laboratory
Associate Professor of Radiology
Associate Professor of Biomedical Engineering
Dept. of Radiology and Dept. of Biomedical Engineering
University of Rochester, New York
601 Elmwood Avenue
Rochester, NY 14642-8648, U.S.A.
Phone: (585) 613-2399
axel_wismueller@urmc.rochester.edu
e-mail: axel@wismueller.de

June 13, 2011

Abstract

In this paper, we present a novel computational framework for nonlinear dimensionality reduction which is specifically suited to process large data sets: the Exploratory Inspection Machine (XIM). XIM introduces a conceptual crosslink between hitherto separate domains of machine learning, namely topographic vector quantization and divergence-based neighbor embedding approaches. There are three ways to conceptualize XIM, namely (i) as the inversion of the Exploratory Observation Machine (XOM) and its variants, such as Neighbor Embedding XOM (NE-XOM), (ii) as a powerful optimization scheme for divergence-based neighbor embedding cost functions inspired by Stochastic Neighbor Embedding (SNE) and its variants, such as t-distributed SNE (t-SNE), and (iii) as an extension of topographic vector quantization methods, such as the Self-Organizing Map (SOM). By preserving both global and local data structure, XIM combines the virtues of classical and advanced recent embedding methods. It permits direct visualization of large data collections without the need for prior data reduction. Finally, XIM can contribute to many application domains of data analysis and visualization important throughout the sciences and engineering, such as pattern matching, constrained incremental learning, data clustering, and the analysis of non-metric dissimilarity data.

1 Motivation

In this paper, we present a novel computational framework for nonlinear dimensionality reduction which is specifically suited to process large data sets: the Exploratory Inspection Machine (XIM). The central idea of XIM is to invert our recently proposed variant of the Exploratory Observation Machine (XOM) [1], [2], [3], namely the Neighbor Embedding XOM (NE-XOM) algorithm as proposed in [4] by systematically exchanging the roles of ordering and exploration spaces in topographic vector quantization mappings. In contrast to NE-XOM, this is accomplished by calculating the derivatives of divergence-based topographic vector quantization cost functions *with respect to distance measures in the high-dimensional data space* rather than derivatives *with respect to distance measures in the low-dimensional embedding space*.

As an important consequence, we can thus introduce successful approaches to alleviate the so-called ‘crowding phenomenon’ as described by [5] into classical topographic vector quantization schemes, in which, in contrast to XOM and its variants, data items are associated with the *exploration* space and not with the *ordering* space of topographic mappings. The resulting learning rules for heavy-tailed neighborhood functions in the low-dimensional embedding space are significantly different from corresponding update rules for NE-XOM and its variants for other divergences.

XIM extends classical topographic vector quantization methods in that its cost function does not only optimize ‘trustworthiness’ in the sense of [6], i.e. small distances in the embedding space correspond to small distances in the data space: By introducing repulsive forces into the learning rules, they optimize ‘continuity’ as well, such that small distances in the data space are represented by small distances in the embedding space. This is exactly opposite to NE-XOM, in which divergence-based cost functions introduce repulsive forces that help to optimize ‘trustworthiness’, whereas the original XOM cost function predominantly optimizes ‘continuity’.

In the light of these considerations, there are three ways to conceptualize XIM as a gen-

eral computational framework for nonlinear dimensionality reduction, namely (i) as the systematic inversion of the Exploratory Observation Machine (XOM) [1], [3] and its variants, such as Neighbor Embedding XOM (NE-XOM) [4], (ii) as a powerful optimization scheme for divergence-based dimensionality reduction cost functions inspired by Stochastic Neighbor Embedding (SNE) [7] and its variants, such as t-distributed SNE (t-SNE) [5], and (iii) as an extension of topographic vector quantization methods, such as the Self-Organizing Map (SOM) [8], where XIM introduces the option to explicitly optimize ‘continuity’. Thus, XIM introduces novel conceptual crosslinks between hitherto separate domains of machine learning.

In the remainder of this paper, we first motivate XIM in the context of topographic vector quantization. We then derive its learning rule from a cost function which is based on the generalized Kullback-Leibler divergence between neighborhood functions in the data and embedding spaces. We further introduce the t-distributed and Cauchy-Lorentz distributed XIM (t-XIM and c-XIM) variants which help to address the crowding phenomenon in nonlinear dimensionality reduction. After specifying technical details of how to calculate the XIM learning rule by computing the derivative of its cost function, we discuss algorithmic properties and variants of XIM, including batch computation and XIM for non-metric data. After presenting experimental results, we finally extend XIM beyond the Kullback-Leibler divergence by deriving XIM learning rules for various other divergence measures known from the mathematical literature.

2 Topographic Mapping

We first motivate the Exploratory Inspection Machine (XIM) in the context of topographic vector quantization techniques. In their simplest form, these algorithms may be seen as methods that map a finite number N of data points $\mathbf{x}_i \in \mathbb{R}^D, i \in \{1, \dots, N\}$ in an exploration space E to low-dimensional image points $\mathbf{y}_i \in \mathbb{R}^d$ in an ordering space O . The assignment is $\mathbf{x}_i \mapsto \mathbf{y}_i$ and typically $d \ll D$, e.g. $d = 2, 3$ for visualization purposes. The mapping $\mathbf{x}_i \mapsto \mathbf{y}_i$ is not calculated directly, rather the data points \mathbf{x}_i are represented by so-called ‘prototypes’

$\mathbf{w}_j \in E \subset \mathbb{R}^D, j \in \{1, \dots, M\}$ that are mapped to target points $\mathbf{r}_j \in O \subset \mathbb{R}^d$, so-called ‘nodes’. These nodes are priorly chosen as a structure hypothesis in the observation space O . Reasonable choices are the location of nodes on a regular lattice structure, e.g. a rectangular or hexagonal grid, but no specific limitations apply, i.e. arbitrary geometrical and even random arrangements may be considered.

Like in topographic vector quantizers, the goal of XIM is to find prototype positions \mathbf{w}_j in such a way that the tuples $(\mathbf{w}_j, \mathbf{r}_j) \in E \times O$ represent pairs of reference points for defining a mapping Ψ which allows the user to find a low-dimensional target point \mathbf{y}_i for each data point \mathbf{x}_i . Such an explicit mapping based on computed pairs of reference points can be accomplished by various additional interpolation or approximation procedures, as shown later.

Each prototype \mathbf{w}_j defines a ‘receptive field’ by a decomposition of E according to some optimality rule for mapping points to prototypes, e.g. by a criterion of minimal distance to a specific prototype. Let the spaces E and O be equipped with distance measures $d_E(\mathbf{x}, \mathbf{w}_j)$ and $d_O(\mathbf{r}_k, \mathbf{r}_l)$. A simple way to define a best-match node \mathbf{r}^* for a given data vector \mathbf{x} is to compute

$$\mathbf{r}^*(\mathbf{x}) = \mathbf{r}_{\phi(\mathbf{x})} = \Psi(\mathbf{w}_{\phi(\mathbf{x})}), \quad (1)$$

where $\phi(\mathbf{x})$ is determined according to the minimal distance criterion

$$d_E(\mathbf{x}, \mathbf{w}_{\phi(\mathbf{x})}) = \min_j (d_E(\mathbf{x}, \mathbf{w}_j)). \quad (2)$$

In a simple iterative approach, learning of prototypes \mathbf{w}_j is accomplished by the online adaptation rule

$$\mathbf{w}_j \leftarrow \mathbf{w}_j - \epsilon h_\sigma(d_O(\mathbf{r}^*, \mathbf{r}_j)) \frac{\partial d_E(\mathbf{x}, \mathbf{w}_j)}{\partial \mathbf{w}_j} \quad (3)$$

where $\epsilon > 0$ denotes a learning rate, d_O and d_E distances in the ordering and exploration spaces, respectively, e.g. squared Euclidean, and $h_\sigma(t)$ a neighborhood cooperativity, which is typically chosen as a monotonically decreasing function, e.g. a Gaussian

$$h_\sigma(t) = \exp\left(-\frac{t}{2\sigma^2}\right), \quad \sigma > 0. \quad (4)$$

However, it should be noted that the distance measures $d_E(\mathbf{x}, \mathbf{w}_j)$ and $d_O(\mathbf{r}_k, \mathbf{r}_l)$ are not confined to squared Euclidean: arbitrary, even non-metric distance measures may be applied. Likewise, the neighborhood cooperativity h is not restricted to Gaussians: other choices, such as heavy-tailed distributions may provide advantages, as discussed below.

3 The Exploratory Inspection Machine (XIM)

It can be shown that for a continuous distribution of inputs, the learning rule (3) cannot be expressed as the derivative of a cost function [9]. According to [10] one can circumvent this problem by replacing the minimal distance criterion (2) by adopting a modified best-match node definition

$$\mathbf{r}^*(\mathbf{x}) = \mathbf{r}_{\phi(\mathbf{x})} = \Psi(\mathbf{w}_{\phi(\mathbf{x})}), \quad (5)$$

where $\phi(\mathbf{x})$ is determined according to

$$\sum_j h_\sigma(d_O(\mathbf{r}_{\phi(\mathbf{x})}, \mathbf{r}_j)) d_E(\mathbf{x}, \mathbf{w}_j) = \min_k \left(\sum_j h_\sigma(d_O(\mathbf{r}_k, \mathbf{r}_j)) d_E(\mathbf{x}, \mathbf{w}_j) \right) \quad (6)$$

This leads to the cost function

$$E' \propto \int \sum_i \delta_{\mathbf{r}^*(\mathbf{x}), \mathbf{r}_i} \sum_{j=1}^M h_\sigma(d_O(\mathbf{r}^*(\mathbf{x}), \mathbf{r}_j)) d_E(\mathbf{x}, \mathbf{w}_j) p(\mathbf{x}) d\mathbf{x}, \quad (7)$$

where δ denotes the Kronecker delta. The derivative of (7) with respect to \mathbf{w}_j using the best-match definition (6) yields the learning rule (3).

Given these settings, we propose to augment the cost function (7) by an additional term in order to combine fast sequential online learning known from topographic mapping, such as in learning rule (3), and principled direct divergence optimization approaches, such as in SNE [7]. To this end, by means of cost function (7) we can define new learning rules based on the generalized Kullback-Leibler divergence for not normalized positive measures p and q with $0 \leq p, q \leq 1$

$$E_{\text{GKL}}(p||q) = \int [p(\mathbf{x}) \log \frac{p(\mathbf{x})}{q(\mathbf{x})}] d\mathbf{x} - \int [p(\mathbf{x}) - q(\mathbf{x})] d\mathbf{x} \quad (8)$$

We use the cooperativity functions $h_\sigma(d_O(\mathbf{r}_i, \mathbf{r}_j))$ and $g_\gamma(d_E(\mathbf{x}, \mathbf{w}_j))$ as positive measures, in the following abbreviated by h_σ^{ij} and g_γ^j .

Inspired by t-SNE [5], neighborhood functions h_σ^{ij} of the low-dimensional embedding space O can be chosen as a heavy-tailed distribution, e.g. a Student t or Cauchy-Lorentzian, in order to alleviate the so-called ‘crowding problem’ which usually occurs in high dimensions related to the curse of dimensionality. It is important to notice that in XIM the neighborhood function of the embedding space refers to the *ordering space* O of the topographic mapping. This marks a fundamental difference when compared to XOM [11] and its variant NE-XOM [4], in which the neighborhood function of the embedding space refers to the *exploration space* E of the topographic mapping.

Based on these settings, we obtain a novel cost function

$$E_{\text{XIM}} \propto \int \sum_i \delta_{\mathbf{r}_{\text{GKL}}^*(\mathbf{x}), \mathbf{r}_i} \sum_j \left[h_\sigma^{ij} \ln \left(\frac{h_\sigma^{ij}}{g_\gamma^j} \right) - h_\sigma^{ij} + g_\gamma^j \right] p(\mathbf{x}) d\mathbf{x} \quad (9)$$

where the best-match node $\mathbf{r}^*(\mathbf{x})$ for data point \mathbf{x} is defined such that

$$\sum_j \left[h_\sigma^{ij} \ln \left(\frac{h_\sigma^{ij}}{g_\gamma^j} \right) - h_\sigma^{ij} + g_\gamma^j \right] \text{ is minimum.} \quad (10)$$

The derivative of this cost function with respect to prototypes \mathbf{w}_j is explained in section 4. It yields the following online learning update rule for a given data vector \mathbf{x} :

$$\mathbf{w}_j \leftarrow \mathbf{w}_j - \epsilon \Delta \mathbf{w}_j \quad (11)$$

with

$$\Delta \mathbf{w}_j = \frac{\partial g_\gamma^j}{\partial \mathbf{w}_j} \left(1 - \frac{h_\sigma^{ij}}{g_\gamma^j} \right). \quad (12)$$

For the special case of a Gaussian neighborhood function in the high-dimensional data space E , i.e. $g_\gamma^j = \exp \left(-\frac{d_E(\mathbf{x}, \mathbf{w}_j)}{2\gamma^2} \right)$, $\gamma > 0$, the learning rule (12) yields

$$\Delta \mathbf{w}_j = \frac{1}{2\gamma^2} \left(h_\sigma^{*j} - g_\gamma^j \right) \frac{\partial d_E(\mathbf{x}, \mathbf{w}_j)}{\partial \mathbf{w}_j}, \quad (13)$$

where h_σ^{*j} may be chosen as a monotonically decreasing function, e.g.

(i) as a Gaussian

$$h_{\sigma}^{*j} = \exp \left(-\frac{d_O(\mathbf{r}^*, \mathbf{r}_j)}{2\sigma^2} \right) \quad (\text{XIM}), \quad (14)$$

(ii) as a t-distribution

$$h_{\sigma}^{*j} = \left(1 + \frac{1}{\sigma} d_O(\mathbf{r}^*, \mathbf{r}_j) \right)^{-\frac{\sigma+1}{2}} \quad (\text{t-XIM}), \quad (15)$$

(iii) as a Cauchy-Lorentz distribution,

$$h_{\sigma}^{*j} = \left(1 + \frac{1}{\sigma^2} d_O(\mathbf{r}^*, \mathbf{r}_j) \right)^{-1} \quad (\text{c-XIM}), \quad (16)$$

where we call the resulting algorithms XIM, t-XIM, and c-XIM, respectively. Obviously (iii) can be seen as a special case of (ii).

For the special choice of a squared Euclidean metric in E , i.e. $d_E(\mathbf{x}, \mathbf{w}_j) = (\mathbf{x} - \mathbf{w}_j)^2$, the learning rule (13) yields

$$\Delta \mathbf{w}_j = -\frac{1}{\gamma^2} \left(h_{\sigma}^{*j} - \exp \left(-\frac{(\mathbf{x} - \mathbf{w}_j)^2}{2\gamma^2} \right) \right) (\mathbf{x} - \mathbf{w}_j). \quad (17)$$

4 Derivative of the XIM Cost Function

With the abbreviation $h_{\sigma}^{ij} = h_{\sigma}(d_O(\mathbf{r}^i, \mathbf{r}^j))$ and $g_{\gamma}^j = g_{\gamma}(d_E(\mathbf{x}^i, \mathbf{w}^j))$, we write the derivative of the cost function equation (9) with respect to a prototype vector \mathbf{w}^k :

$$\begin{aligned} \frac{\partial E_{\text{XIM}}}{\partial \mathbf{w}^k} &\propto \int \sum_i \frac{\partial \delta_{\Psi^{\text{GKL}}(\mathbf{x}), \mathbf{r}^i}}{\partial \mathbf{w}^k} \sum_j \left[h_{\sigma}^{ij} \ln \left(\frac{h_{\sigma}^{ij}}{g_{\gamma}^j} \right) - h_{\sigma}^{ij} + g_{\gamma}^j \right] p(\mathbf{x}) d\mathbf{x} \\ &+ \int \sum_i \delta_{\Psi^{\text{GKL}}(\mathbf{x}), \mathbf{r}^i} \frac{\partial g_{\gamma}^k}{\partial \mathbf{w}^k} \left(1 - \frac{h_{\sigma}^{ik}}{g_{\gamma}^k} \right) p(\mathbf{x}) d\mathbf{x}, \end{aligned} \quad (18)$$

with $\mathbf{r}^{*\text{GKL}}(\mathbf{x}) = \mathbf{r}^*(\mathbf{x})$ as defined in equation (10). The latter term yields the learning rule. The first term vanishes, as can be seen as follows: We use the shorthand notation

$$\Phi^N(\mathbf{r}^i, \mathbf{x}) = \sum_j \left[h_{\sigma}^{ij} \ln \left(\frac{h_{\sigma}^{ij}}{g_{\gamma}^j} \right) - h_{\sigma}^{ij} + g_{\gamma}^j \right]. \quad (19)$$

Then the best-match node can be expressed as

$$\delta_{\mathbf{r}^* \text{GKL}(\mathbf{x}), \mathbf{r}^i} = H \left(\sum_k H(\Phi^N(\mathbf{r}^i, \mathbf{x}) - \Phi^N(\mathbf{r}^k, \mathbf{x})) - n + \frac{1}{2} \right). \quad (20)$$

Hence the additional first term of equation (18) vanishes, because of the following:

$$\begin{aligned} & \int \sum_i \frac{\partial \delta_{\mathbf{r}^* \text{GKL}(\mathbf{x}), \mathbf{r}^i}}{\partial \mathbf{w}^k} \Phi^N(\mathbf{r}^i, \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= \int \sum_i \delta \left(\sum_l H(\Phi^N(\mathbf{r}^i, \mathbf{x}) - \Phi^N(\mathbf{r}^l, \mathbf{x})) - n + \frac{1}{2} \right) \\ & \quad \cdot \sum_l \delta(\Phi^N(\mathbf{r}^i, \mathbf{x}) - \Phi^N(\mathbf{r}^l, \mathbf{x})) \cdot \left[(g_{\gamma}^k - h_{\sigma}^{ik}) - (g_{\gamma}^k - h_{\sigma}^{lk}) \right] \frac{\partial g_{\gamma}^k}{\partial \mathbf{w}^k} \end{aligned} \quad (21)$$

$$\begin{aligned} & \sum_j \left[h_{\sigma}^{ij} \ln \left(\frac{h_{\sigma}^{ij}}{g_{\gamma}^j} \right) - h_{\sigma}^{ij} + g_{\gamma}^j \right] p(\mathbf{x}) d\mathbf{x} \\ &= \int \sum_{ilj} \delta \left(\sum_{l'} H(\Phi^N(\mathbf{r}^i, \mathbf{x}) - \Phi^N(\mathbf{r}^{l'}, \mathbf{x})) - n + \frac{1}{2} \right) \\ & \quad \cdot \delta(\Phi^N(\mathbf{r}^i, \mathbf{x}) - \Phi^N(\mathbf{r}^l, \mathbf{x})) \cdot (g_{\gamma}^k - h_{\sigma}^{ik}) \frac{\partial g_{\gamma}^k}{\partial \mathbf{w}^k} \\ & \quad \cdot \left[h_{\sigma}^{ij} \ln \left(\frac{h_{\sigma}^{ij}}{g_{\gamma}^j} \right) - h_{\sigma}^{ij} + g_{\gamma}^j \right] p(\mathbf{x}) d\mathbf{x} \\ & \quad - \int \sum_{ilj} \delta \left(\sum_{l'} H(\Phi^N(\mathbf{r}^i, \mathbf{x}) - \Phi^N(\mathbf{r}^{l'}, \mathbf{x})) - n + \frac{1}{2} \right) \\ & \quad \cdot \delta(\Phi^N(\mathbf{r}^l, \mathbf{x}) - \Phi^N(\mathbf{r}^i, \mathbf{x})) \cdot (g_{\gamma}^k - h_{\sigma}^{ik}) \frac{\partial g_{\gamma}^k}{\partial \mathbf{w}^k} \\ & \quad \cdot \left[h_{\sigma}^{lj} \ln \left(\frac{h_{\sigma}^{lj}}{g_{\gamma}^j} \right) - h_{\sigma}^{lj} + g_{\gamma}^j \right] p(\mathbf{x}) d\mathbf{x} \end{aligned} \quad (22)$$

$$\begin{aligned}
& \sum_j \left[h_{\sigma}^{ij} \ln \left(\frac{h_{\sigma}^{ij}}{g_{\gamma}^j} \right) - h_{\sigma}^{ij} + g_{\gamma}^j \right] p(\mathbf{x}) d\mathbf{x} \\
&= \int \sum_{il} \delta \left(\sum_{l'} H(\Phi^N(\mathbf{r}^i, \mathbf{x}) - \Phi^N(\mathbf{r}^{l'}, \mathbf{x})) - n + \frac{1}{2} \right) \\
&\quad \cdot \delta(\Phi^N(\mathbf{r}^i, \mathbf{x}) - \Phi^N(\mathbf{r}^l, \mathbf{x})) \cdot (g_{\gamma}^k - h_{\sigma}^{ik}) \frac{\partial g_{\gamma}^k}{\partial \mathbf{w}^k} \cdot \Phi^N(\mathbf{r}^i, \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\
&\quad - \int \sum_{il} \delta \left(\sum_{l'} H(\Phi^N(\mathbf{r}^l, \mathbf{x}) - \Phi^N(\mathbf{r}^{l'}, \mathbf{x})) - n + \frac{1}{2} \right) \\
&\quad \cdot \delta(\Phi^N(\mathbf{r}^l, \mathbf{x}) - \Phi^N(\mathbf{r}^i, \mathbf{x})) \cdot (g_{\gamma}^k - h_{\sigma}^{ik}) \frac{\partial g_{\gamma}^k}{\partial \mathbf{w}^k} \cdot \Phi^N(\mathbf{r}^l, \mathbf{x}) p(\mathbf{x}) d\mathbf{x} = 0, \tag{23}
\end{aligned}$$

because of the symmetry of δ and the fact that δ is non-vanishing only if $\Phi^N(\mathbf{r}^l, \mathbf{x}) = \Phi^N(\mathbf{r}^i, \mathbf{x})$.

5 Summary of the XIM Algorithm

In summary, the XIM algorithm encompasses the following steps:

Exploratory Inspection Machine (XIM)

Input: Data vectors \mathbf{x} or pairwise dissimilarities $d_E(\mathbf{x}_i, \mathbf{x}_k)$ (see section 7)

Initialize: Number of iterations t_{\max} , learning rate ϵ , cooperativity parameters σ, γ , structure hypothesis \mathbf{r}_j , initial prototypes \mathbf{w}_j

For $t \leq t_{\max}$ **do**

begin

1. Randomly draw a data vector \mathbf{x} .
2. Find best-match node using equation (10).
3. Compute neighborhood cooperativity h_σ (may be pre-computed and retrieved from a look-up table).
4. Compute neighborhood cooperativity g_γ
5. Update prototypes according to equations (11) and (12), e.g. using equations (13) – (17).

end

Output: Prototypes \mathbf{w}_j

Optional: Compute explicit mapping by applying approximation or interpolation schemes to explicitly calculate a low-dimensional target point \mathbf{y} for each data point \mathbf{x} , see section 6.

6 Algorithmic Properties and Implementation Issues

XIM, unlike SNE and many other embedding algorithms, exhibits the interesting property that it allows to impose a prior structure on the projection space, which is a property that can also be found in XOM and SOM. Unlike SNE and many other data visualization techniques which

exhibit a computational and memory complexity that is quadratic in the number of data points, the complexity of XIM can be easily controlled by the structure hypothesis definition, i.e. the node grid in the observation space, and is linear with the number of data points and the number of prototypes. When implementing XIM, the following issues may be considered:

Best-match definition. Although from a theoretical point of view, the best-match definition should be made according to equation (10), we found that using the much simpler minimal distance approach according to equation (6) yields comparable results in practical computer simulations.

Weighting of attractive and repulsive forces. It should be noted that the factor $\frac{1}{\gamma^2}$ in the update rules (13) and (17) can be seen as a quantity that determines the stepsize of a gradient descent step on the cost function, which can be treated as a free parameter similarly as the learning parameter ϵ in the online learning rule (11). We found it useful to omit the factor $\frac{1}{\gamma^2}$ and control the stepsize of gradient descent by annealing of the learning parameter ϵ only and to introduce a relative weighting of attractive and repulsive forces in learning rule (17), i.e.

$$\Delta \mathbf{w}_j = - \left((1 - \eta) h_{\sigma}^{*j} - \eta \exp \left(- \frac{(\mathbf{x} - \mathbf{w}_j)^2}{2\gamma^2} \right) \right) (\mathbf{x} - \mathbf{w}_j), \quad (24)$$

where the method proved robust for a wide range of values for $\eta \in [0.1, 0.5]$.

Similarly as in other topographic mapping approaches, the parameters ϵ , σ , and γ may be adapted using annealing schemes, e.g. by using an exponential decay

$$\kappa(t) = \kappa(t_1) \left(\frac{\kappa(t_{\max})}{\kappa(t_1)} \right)^{\frac{t}{t_{\max}}} \quad (25)$$

where $\kappa := \epsilon$, $\kappa := \sigma$, or $\kappa := \gamma$, respectively, and $t \in [t_1, t_{\max}]$ denotes the iteration step.

As known from other algorithms, such as SNE or XOM, it can be useful to adapt the variance of the neighborhood cooperativity in the data space to the local ‘density’ of neighbors, i.e. to use a different γ_i for each data sample \mathbf{x}_i so that an α -ball of variance γ_i would contain a fixed number k of neighbors. The rationale of this would be to ensure that data samples in less dense

regions are adequately represented in the calculation of the embedding. Instead of annealing γ directly, this number k may be annealed, e.g. according to equation (25) by setting $\kappa := k$. An alternative approach is to find appropriate γ_i by using the ‘perplexity’ approach proposed for SNE in [7].

Explicit mapping of data to target points. As the output of XIM is the location of prototypes \mathbf{w}_j in the exploration space E , there is no explicit mapping of each data vector onto its target points. This marks a fundamental difference between XIM and XOM: As in XOM, each data vector has its own prototype, the final prototype positions already represent the low-dimensional embeddings for each data point. This does not hold for XIM. However, an explicit mapping can be accomplished easily by using the tuples $(\mathbf{w}_j, \mathbf{r}_j) \in E \times O$ as pairs of reference points for defining a mapping Ψ which allows the user to find a low-dimensional target point \mathbf{y}_i for each data point \mathbf{x}_i . Such an explicit mapping based on computed pairs of reference points can be calculated by a wide range of interpolation or approximation procedures. Here, we specifically mention three approaches that have been successfully applied in similar contexts, namely Shepard’s interpolation [12], generalized radial basis functions softmax interpolation, e.g. [13], and the application of supervised learning methods. For the latter approach, $(\mathbf{w}_j, \mathbf{r}_j)$ pairs are used as labeled examples for models of supervised learning, such as feed-forward neural networks or other function approximators. For technical details of how to use these methods in a related context, we refer to our previous work [13]. It should be noted that such explicit mapping approaches can also be used for ‘out-of-sample’ extension, i.e. for finding embeddings of new data points that have not previously been used for XIM training.

7 Batch XIM and XIM for Non-Metric Data

Batch XIM algorithms can easily be derived from analyzing the stationary states of learning rules (12) – (13), where we expect the expectation values for these updates over all data points to be zero. For example, for learning rule (13), we expect the following condition to be fulfilled

in the stationary state

$$\mathcal{E}_i \left(f_{\mathbf{r}^*(\mathbf{x}_i), \mathbf{r}_j} \frac{\partial d_E(\mathbf{x}_i, \mathbf{w}_j)}{\partial \mathbf{w}_j} \right) = 0 \quad \forall j \in \{1, \dots, M\}, \quad (26)$$

where \mathcal{E}_i denotes the expectation value over all data points \mathbf{x}_i and $f_{\mathbf{r}^*(\mathbf{x}_i), \mathbf{r}_j} = h_{\sigma}^{*j} - g_{\gamma}^j$. For the squared Euclidean metric, we obtain the batch XIM fixed-point iteration

$$\mathbf{w}_j = \frac{\sum_i f_{\mathbf{r}^*(\mathbf{x}_i), \mathbf{r}_j} \mathbf{x}_i}{\sum_i f_{\mathbf{r}^*(\mathbf{x}_i), \mathbf{r}_j}}. \quad (27)$$

In analogy to equations (6) – (8) in [14], even simpler and faster batch XIM methods can easily be obtained by introducing Voronoi sets of the exploration space according to the best-match criterion (10). For practical implementations, however, the simple minimal distance criterion (5) may be used.

Finally, we introduce **XIM for non-metric data** by combining the batch XIM with the concept of the generalized median. This can be accomplished in full analogy to SOM for nonvectorial data according to [15]. However, a fundamental difference is that in forming the sum of distances for computing the generalized median, the contents of the node-specific data sublists within a neighborhood set of grid nodes should be weighted by the above cooperativity $f_{\mathbf{r}^*(\mathbf{x}_i), \mathbf{r}_j}$ rather than h_{σ}^{*j} alone as in [15]. Likewise, the best-match definition (10) should be used instead of a simple minimal distance criterion (5).

8 Experiments

To prove the applicability of XIM, we present results on real-world data, see Figs. 1 and 2. The data consists of 147 feature vectors in a 79-dimensional space encoding gene expression profiles obtained from microarray experiments.

Fig. 1 shows visualization results obtained from structure-preserving dimensionality reduction of gene expression profiles related to ribosomal metabolism, as a detailed visualization of a subset included in the genome-wide expression data taken from Eisen et al. [16]. The figure illustrates the exploratory analysis of the 147 genes labeled as ‘5’ (22 genes) and ‘8’ (125

genes) according to the cluster assignment by Eisen et al. [16]. Besides several genes involved in respiration, cluster ‘5’ (blue) contains genes related to mitochondrial ribosomal metabolism, whereas cluster ‘8’ (red) is dominated by genes encoding ribosomal proteins and other proteins involved in translation, such as initiation and elongation factors, and a tRNA synthetase. The data has been described in [17].

In the c-XIM genome map of Fig. 1A using a square grid of 30×30 nodes in the ordering space, it is clearly visible at first glance that the data consists of two distinct clusters. Comparison with the functional annotation known for these genes [18] reveals that the map overtly separates expression profiles related to mitochondrial and to extramitochondrial ribosomal metabolism. Fig. 1B shows a data representation obtained by a Self-Organizing Map (SOM) trained on the same data also using a square grid of 30×30 nodes in the ordering space. As can be clearly seen in the figure, SOM *cannot* achieve a satisfactory cluster separation in the mapping result as provided by c-XIM in Fig. 1A: Although the genes related to mitochondrial and to extramitochondrial ribosomal metabolism are collocated on the map, the distinct cluster structure underlying the data remains invisible, if the color coding is omitted. The result obtained by the Exploratory Observation Machine (XOM) in Fig. 1C as well as the mapping result obtained by Principal Component Analysis (PCA) in Fig. 2, however, can properly recover the underlying cluster structure. Note that the result obtained by XIM using Gaussian neighborhood functions in the ordering space (Fig. 1D) resembles the SOM result in that a clear cluster separation cannot be observed. The comparison of Figs. 1A and D thus demonstrate the benefit of introducing heavy-tailed neighborhood functions in the ordering space for this example.

A quantitative comparison of several quality measures for the results obtained in the data set of Figs. 1 and 2 is presented in Tab. 1.

Table 1: Comparative evaluation of quantitative embedding quality measures as known from the literature for the results obtained for dimensionality reduction of genome expression profiles related to ribosomal metabolism in Figs. 1 and 2, namely Sammon’s error [19], Spearman’s ρ , as well as ‘trustworthiness’ and ‘continuity’ as defined by [6]. The table encompasses mean values and standard deviations obtained for 10 runs of each method, where in each run 140 out of 147 data points (approx. 95% of the data) are randomly sub-sampled, i.e. embeddings and corresponding quality measures are computed for each of the 10 sub-sampled data sets. Trustworthiness and continuity values are computed as mean values obtained for $k = 1, \dots, 50$ neighbors. The free parameters of all the methods examined in the comparison (except PCA) were optimized to obtain the best results with regard to the respective quality measure. Explicit mappings for c-XIM and SOM were computed using Shepard’s interpolation [12]. Note that in comparison to SOM, c-XIM yields competitive results for several quality measures.

Method	Sammon	Spearman’s ρ	Trustworthiness	Continuity
SOM	0.18 (0.02)	0.50 (0.05)	0.84 (0.02)	0.85 (0.01)
c-XIM	0.17 (0.02)	0.59 (0.09)	0.87 (0.02)	0.86 (0.02)
XOM	0.19 (0.03)	0.88 (0.02)	0.86 (0.02)	0.90 (0.02)
PCA	0.18 (0.01)	0.86 (0.01)	0.85 (0.01)	0.89 (0.01)

9 XIM Using Other Divergences

XIM, as described so far, has been introduced based on the generalized Kullback-Leibler divergence. However, it is important to emphasize that XIM is not restricted to this specific divergence measure alone, but other divergences can be used easily. However, our work presented here differs substantially from prior approaches to utilize divergences for nonlinear dimensionality reduction. In contrast to SNE [7] and t-SNE [5], divergence-based cost functions for nonlinear dimensionality reduction are *not* optimized *directly*, but via iterative incremental online learning. In contrast to [20], we do not use divergences as distance measures *within* the data or the embedding space, but as a dissimilarity measure *between* these spaces. In contrast to our previous work [4], which applies divergences to XOM, the XIM method presented in this

paper systematically inverts the quoted approaches by introducing divergences in order to derive novel topographic vector quantization learning rules. This is accomplished by calculating the derivatives of divergence-based topographic vector quantization cost functions *with respect to distance measures in the high-dimensional data space* rather than derivatives *with respect to distance measures in the low-dimensional embedding space*, such as in [4] or [21].

As an important consequence, we can thus introduce successful approaches to alleviate the so-called ‘crowding phenomenon’ as described by [5] into classical topographic vector quantization schemes, in which, in contrast to XOM and its variants, data items are associated with the *exploration* space and not with the *ordering* space of topographic mappings. Specifically, the deep difference when compared to the XOM approach becomes evident, when heavy-tailed distributions in the low-dimensional embedding space are utilized to address this issue: The update rules for the resulting t-XIM (and c-XIM) algorithms are significantly different from the t-NE-XOM update rules as specified in [4] for the Kullback-Leibler divergence and in [21] for other divergences. The reason is that for t-XIM and c-XIM, in contrast to NE-XOM and its variants, there is no need to re-compute the derivative of the neighborhood function h of the low-dimensional embedding space with respect to its underlying distance measure when computing the new learning rules. Instead, only h , not g has to be replaced by a heavy-tailed distribution. In other words, we can derive learning rules for XIM by directly adopting derivatives of divergences as known from the mathematical literature and creatively applying them to a completely different context.

10 Extensions of XIM

XIM is likely to be even more useful by slight extensions or in combination with other methods of data analysis and statistical learning. The combination with interpolation and approximation schemes for explicit mapping between data and embedding spaces has already been discussed above. Other straightforward variants of XIM comprise the iterative use of XOM, local or

hierarchical data processing schemes, e.g. by initial vector quantization, and variable choices of batch, growing, or speed-up variants analogous to those described in the literature on topology-preserving mappings.

As noted earlier, there are no principal restrictions to input data and structure hypotheses. For example, they may be subject to a non-Euclidean, e.g. hyperbolic, geometry, or even represent nonmetric dissimilarities. Distances in the observation and exploration spaces may be rescaled dynamically during XIM training. These slight extensions should further enhance the applicability of XIM to many areas of information processing.

Acknowledgements

The author thanks Markus B. Huber and Mahesh B. Nagarjan for performing computer simulations on XIM.

References

- [1] A. Wismüller. Exploratory Morphogenesis (XOM): A Novel Computational Framework for Self-Organization. Ph.D. thesis, Technical University of Munich, Department of Electrical and Computer Engineering, 2006.
- [2] Axel Wismüller. The exploration machine - a novel method for data visualization. In *Lecture Notes in Computer Science. Advances in Self-Organizing Maps*, pages 344–352, 2009.
- [3] A. Wismüller. The Exploration Machine – a novel method for structure-preserving dimensionality reduction. In M. Verleysen, editor, *European Symposium on Artificial Neural Networks (ESANN)*, pages 59–64, Bruges, Belgium, 2009.

- [4] K. Bunte, B. Hammer, T. Villmann, M. Biehl, and A. Wismüller. Neighbor embedding XOM for dimension reduction and visualization. *Neurocomputing*, 74(9):1340 – 1350, 2011.
- [5] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.
- [6] J. Venna. Dimensionality reduction for visual exploration of similarity structures. Ph.D. thesis, Helsinki University of Technology, 2006.
- [7] G. Hinton and S. Roweis. Stochastic neighbor embedding. In *Neural Information Processing Systems 15*, pages 833–840. MIT Press, 2003.
- [8] T. Kohonen. *Self-Organizing Maps*. Springer, Berlin, Heidelberg, New York, 3rd edition, 2001.
- [9] W. Erwin, K. Obermayer, and K. Schulten. Self-organizing maps: ordering, convergence properties, and energy functions. *Biological Cybernetics*, 61:47–55, 1992.
- [10] T. Heskes. Energy functions for self-organizing maps. In E. Oja and S. Kaski, editors, *Kohonen Maps*, pages 303–315. Elsevier, Amsterdam, 1999.
- [11] A. Wismüller. A computational framework for exploratory data analysis in biomedical imaging. In *Proceedings of SPIE*, volume 7260, 2009. doi:10.1117/12.813892.
- [12] D. Shepard. A two-dimensional interpolating function for irregularly spaced data. In *Proc. ACM. nat. Conf.*, pages 517–524, 1968.
- [13] A. Wismüller, F. Vietze, D.R. Dersch, J. Behrends, K. Hahn, and H. Ritter. The deformable feature map – a novel neurocomputing algorithm for adaptive plasticity in pattern analysis. *Neurocomputing*, 48:107–139, 2002.

- [14] T. Kohonen. The self-organizing map. *Neurocomputing*, 21:1–6, 1998.
- [15] T. Kohonen. How to make large self-organizing maps for nonvectorial data. *Neural Networks*, 15:945–952, 2002.
- [16] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998.
- [17] Axel Wismüller. A computational framework for nonlinear dimensionality reduction and clustering. In *Lecture Notes in Computer Science. Advances in Self-Organizing Maps*, pages 334–343, 2009.
- [18] J.M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, and R.K. Mortimer. *Nature*, 387:67–73, 1997.
- [19] J.W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C 18:401–409, 1969.
- [20] T. Villmann and S. Haase. Divergence-based vector quantization. *Neural Networks*, 15:945–952, 2002.
- [21] K. Bunte, T. Haase, M. Biehl, and T. Villmann. Mathematical foundations of self organized neighbor embedding (SONE) for dimension reduction and visualization. In *Machine Learning Reports*, number 03/2010. University of Applied Sciences, Mittweida, 2010.

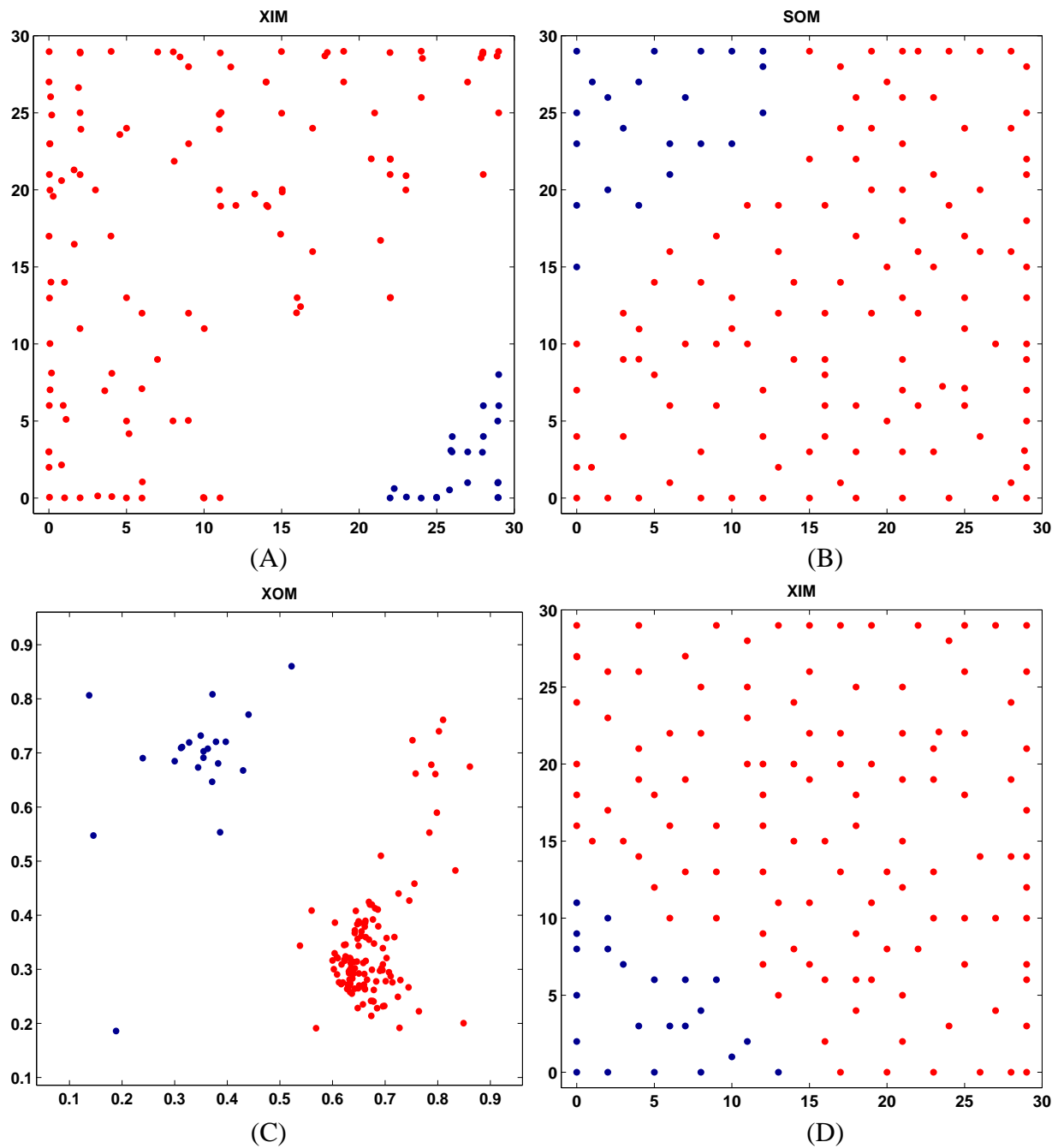


Figure 1: Nonlinear dimensionality reduction of genome expression profiles related to ribosomal metabolism using (A) Cauchy-Lorentz Exploratory Inspection Machine (c-XIM), (B) Self-Organizing Map (SOM), (C) Exploratory Observation Machine (XOM), and (D) XIM with Gaussian neighborhood functions in the ordering space. For explanation, see text.

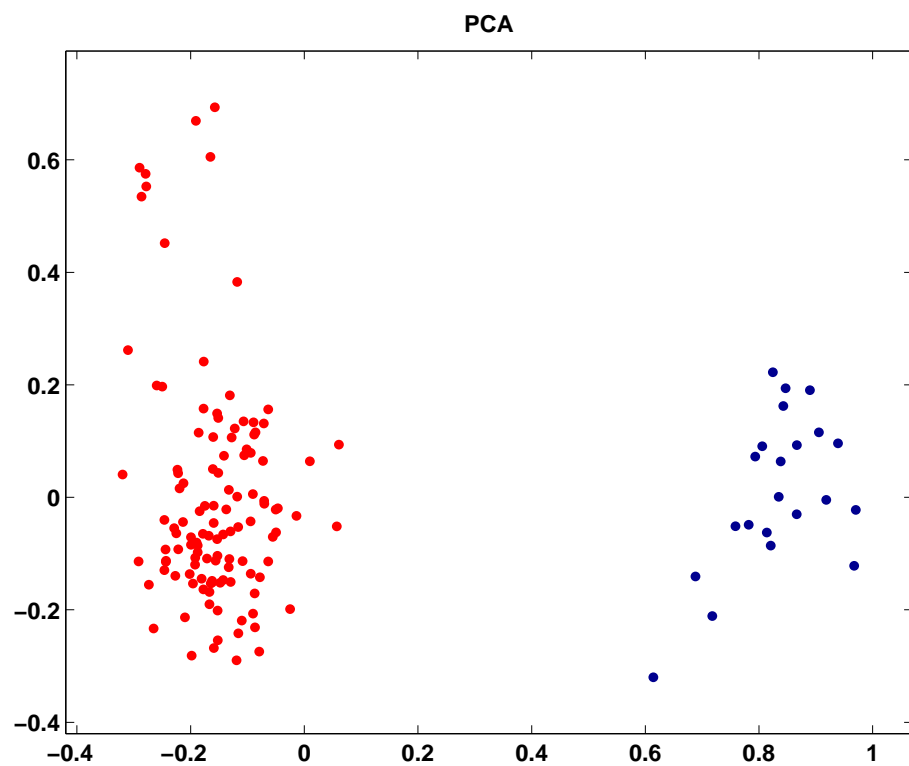


Figure 2: Dimensionality reduction of genome expression profiles related to ribosomal metabolism using Principal Component Analysis (PCA). For explanation, see text.